Project no. 043338

Project acronym: EMERGENCE

Project title: A foundation for Synthetic Biology in Europe

Instrument: NEST Pathfinder

Thematic Priority: Synthetic Biology

D3.1: Document describing the concepts for integrated workflow infrastructure based on the registry
D3.2: Report describing the implementation of software and the integration tools and methods for sequence design and analysis

Due date of deliverable: D3.1: May 2007; D3.2: December 2007
Actual submission date: December 2007

Start date of project: 1.12.2006 Duration: 36 months

Alfonso Valencia
Structural and Computational Biology Programme Spanish National Cancer Research Centre (CNIO) Melchor Fernandez Almagro, 3. E-28029 Madrid

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

On the integration of Bioinformatic Tools and Methods for Synthetic Biology and the Registry of Biological Parts.

This document describes the first steps in the creation of the bioinformatics infrastructure for connecting the MIT 'Registry of Standardized Biological Parts' (http://partsregistry.org/) with the biological databases, repositories and tools.

This linkage was considered a key element for a European IT infrastructure for Synthetic Biology and for the general facilitation of the usage of the information accumulated in the Registry in real experimental biology.

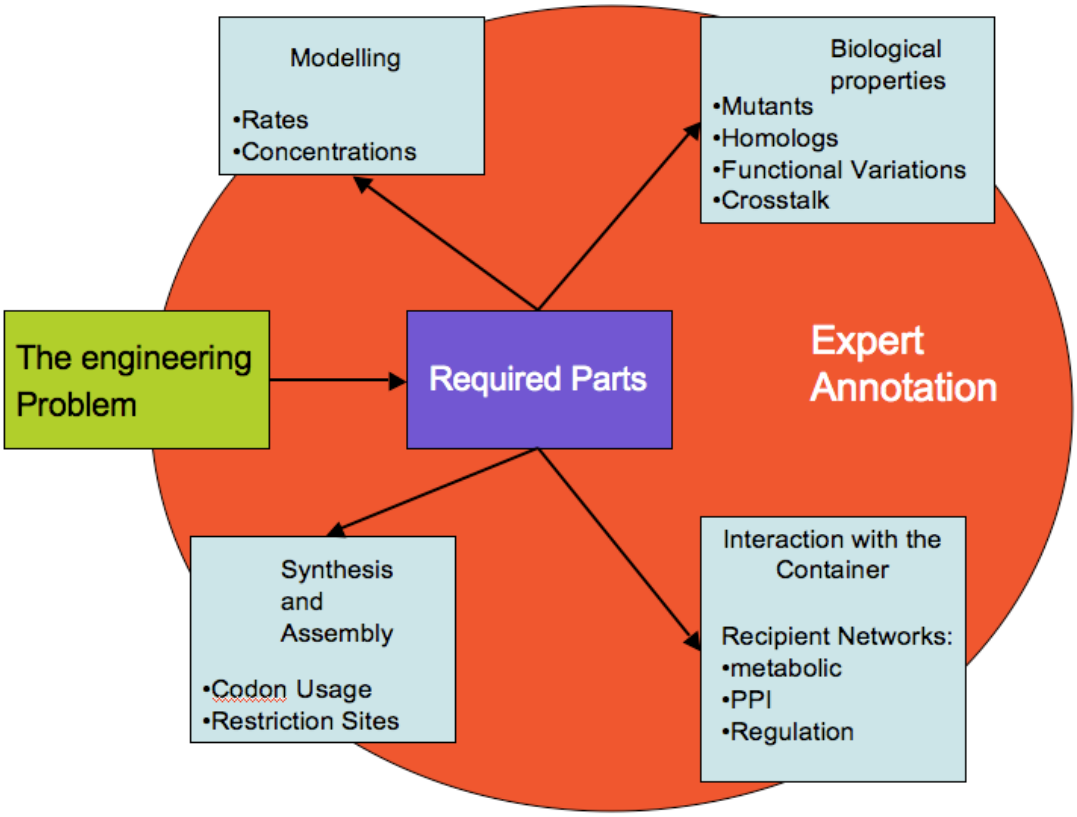
The information provided in the document is based on collaboration with Randy Rettberg and the CNIO group, including the three days hands-on workshop in Madrid (5 and 6/9/2007) and the participation of Ildefonso Cases at the iGEM Jamboree atMIT last November.

The document includes analysis of the organization of the MIT Registry, the update of the more appropriate genomic methods and tools and the first implemented prototypes.

All those developments are state of the art and in a preparatory phase, given the MIT Registry is open to the contribution of a large community of developers (http://partsregistry.org/) and by its nature can be considered in a pre-consolidation phase.

1. The need for bioinformatics of methods in Synthetic Biology: Tools and methods on the genomic/ protein analysis.

In the typical Synthetic Biology scenario after the formulation of the basic design at the general conceptual-engineering level, the implementation is carried out actual biological parts (i.e. ORFs, promoters, etc) that have to be selected from the set of known ones. The full system is then modelled and simulated, and finally synthesized, assembled and integrated in the biological chassis.



The process of selection of the biological parts can be assisted by a number of bioinformatics tools that will inform the Synthetic Biology engineer of what is available, what has been tested, and very importantly of what is the biological context in which these parts are working.

The basic steps in which Bioinformatics tools can be useful are described in the following three points.

1.- For the selection of the biological parts are:

-Identification of available parts in the Registry with the required features (see below)

- Literature and Database mining for identification of properties of the parts, such us enzymatic and kinetic properties and identification of characterized mutants with specific functional alterations.

- Identification of other functional equivalents, either homologues or not, with more suited properties for Synthetic Biology in general (smaller, more specific, etc) or for the biological project in particular.

- Identification of potential functional modifications via mutation, by providing structural information, and in particular information about active or regulatory sites.
- *Orthogonalization*: Identification of interactions with the "chassis" or other parts and prediction of their effects on the designed device, including metabolic, regulatory or protein-protein interactions

2.- Modelling the predicted behaviour of the system/device:

Beside the collection of modelling tools available, bioinformatic tools can also help in filling in some information that is often missing from repositories and can be identify in the literature or custom databases, such as expression, concentration, binding constants, etc. Also tools for computer-assisted design of biological circuits are starting (see second part of the workpackage)

3.- Synthesis and assembly.

In this phase bioinformatics tools can help in several aspects from codon usage, presence of restriction sites and how to remove those producing silent mutations, etc.

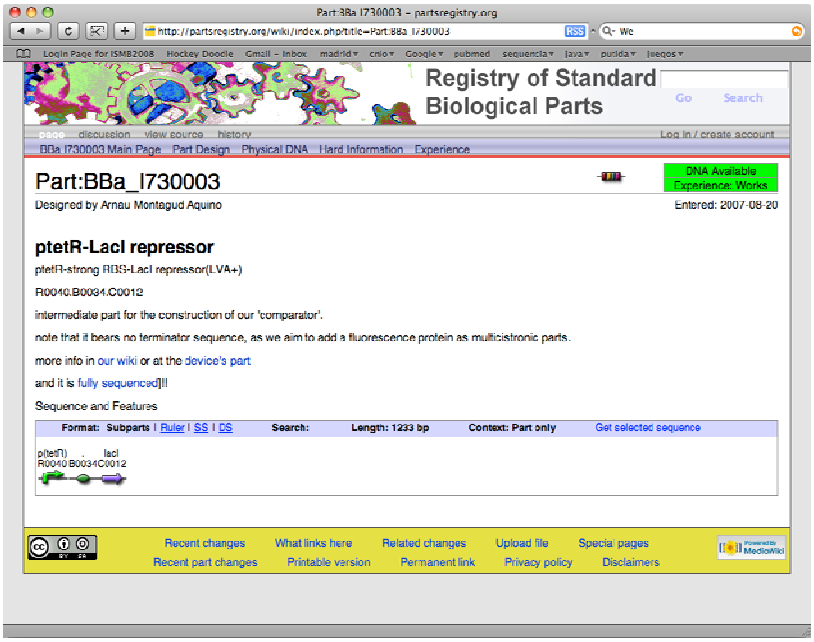
An additional value arise from the possibility of daisy-chain all these applications, creating pipe-lines that will automate, at least part of a laborious work currently done manually.

2. Analysis of the requirement of tools and methods to work in association with the MIT Registry

Since a key concept in Synthetic Biology is the re-utilization and standardization, it is fundamental the creation of catalogues of parts, that can be reused and combined in new devices with more complex and useful functions. The most populated catalogue of biological parts for Synthetic Biology is the MIT Registry of Biological Parts (http://partsregistry.org/), which currently store information about more than 5000 parts.

Currently is a relational database with a web interface that provides access to part descriptions and some sequence analysis tools (including blast) embedded in the web site. Parts are categorized in three main classes: Systems, Devices, and Parts; each

including a number subcategories. Information for each parts varies enormously and relies on that provided by the creator of the part, and include as minimum the DNA sequence, an unique ID, and free-text description and some details about the availability and experience of use. Most of this information is not structured or computer tractable. No ontologies or controlled vocabularies are in use, and not a formal description of parts and relations between parts are available.



The Registry has been extremely useful and fundamental for the development of the Synthetic Biology as an established discipline. However, with increasing number of parts and users and the rising complexity of Synthetic Biology projects, a number of improvements could be beneficial to reach its full potential. The Registry is addressing some of these limitations and is currently under redesigning and improvement, increasing the number of available tools, making some of the information computer readable and reorganizing parts categories (http://partsregistry.org/BioBrick_Part_Program) In parallel, the Registry is encouraging part contributors to better characterize and document their part through a Part Promotion Process based on the well establish system of peer-review (see http://partsregistry.org/Part_Promotion_Process)

Many of these improvements could be fully or partially addressable by bioinformatics means. These are some examples:

Quality Checks:

-Confirm user provided annotation (CDS, promoters, etc)

-Check redundancy and subparts

-Presence of non-allowed restriction sites

-Predict expression/cloning/amplification problems

Cross-Linking with other databases:

-Match part sequence to external databases such as Genebank/EMBLBank,Uniprot, Genome Databases, etc.

Use of Ontologies and Control Vocabularies:

-Normalize vocabulary for elements, as promoters, etc, using standard ontologies (i.e. Sequence Ontology)

Add/Link to Additional Data:

-Variations,

-Homologues,

-Properties (other substrates, inhibitors, Km, etc)

-Associated literature.

3. Study of the technical feasibility of the connectivity of the MIT Registry with the Biological databases.

In order of to be able to benefit from the large set of available Bioinformatics tools, it is necessary to create the methods for exporting the information contained in the Registry in a standard computer-tractable manner. Several alternatives are in common use on other databases in different areas of bioinformatics.

Brief description:

Flat-File Distribution: The simplest option for distribution of data involves packing the database in a custom or defined standard, and then allow users to download it. However a number of problems arise from this model, such the necessity of strict format and versioning control, long inter-releases periods, etc.

SQL: Other databases as Cisred and also Ensembl, provide direct access to their database by using standard query languages as SQL or SPARQL. This solves partly the previous problems since modifications in the database are immediately available. While this option offers complete reading access to the data in an absolutely flexible manner,

the developer has to familiarize with the database schema, something that can be difficult for complex data structures and thus limiting access.

DAS: stands for *Distributed Annotation System*. Based on a Reference Server and Annotation servers that provide information base on entities and coordinates provided by the reference server, allowing the integration of annotation from different sources, as long as they share the same reference server. It is in use in well-established databases as Uniport, InterPro, Pfam, Ensembl, UCSC Genome database, etc. Recent versions allow annotation of not only DNA sequences, but also protein sequences or even Proteins structures, and the distribution of not only annotations but also alignments. REST and XML based, The DAS protocol defined a set of defined queries and responses, and although they cover a most of the usual user request, it does not provide the flexibility of other alternatives, and most notably DAS does not provide search request. There are many implementations of both servers and clients already available, along with libraries that allow rapid implementation of client software.

Webservices/SOAP/BioMoby: web services, and in particular SOAP services allow easy implementation of clients taking advantage of XML base standards. In particular BioMoby services are a breath of services specially developed for bioinformatics Tools, including strong biological entities typing, process control, error control, etc. BioMoby services can be easily assembled in pipelines and these can be designed using graphical interfaces such as “Taverna”.

API: A final alternative is to develop a set of libraries specific for data access. A main advantage is that the data schema is complete hidden so it can evolve and change without disrupting the access to external uses.

Another aspect of the Registry related with the distribution of information is related with the potential coordination of different parts repositories. This issue relates with other problems like synchronization and distribution of part information across different repositories, minimal description of parts, formats, etc. The exploration of these aspects of the Registry, while related, is at this point out of the scope of this report. However, since it will ultimately affect also the way the Registry communicates with analysis and modelling tools, the EMERGENCE action is also participating in this discussion with the presence of the group of Luis Serrano (see the BioBrick Standards Mailing List http://biobricks.org/pipermail/standards_biobricks.org/)

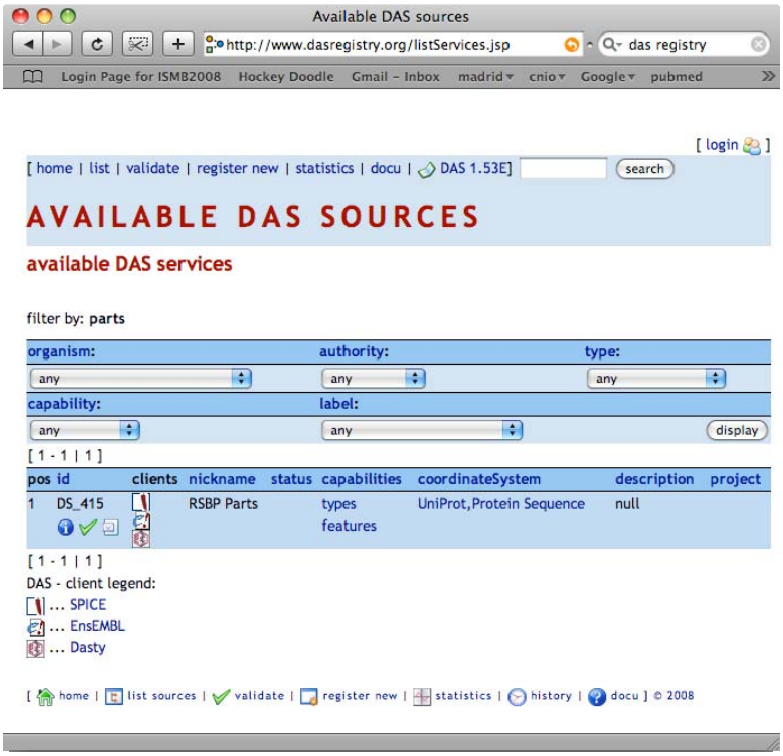
4. Initial implementations to make the Registry accessible to the genomics tools

During the technical workshop in 2007 we discarded the implementation of a direct access to the MIT database due to the complexity of the database structure. Between SOAP/BioMoby services and DAS, we selected DAS as prototype mainly due to the easier nature of its implementation.

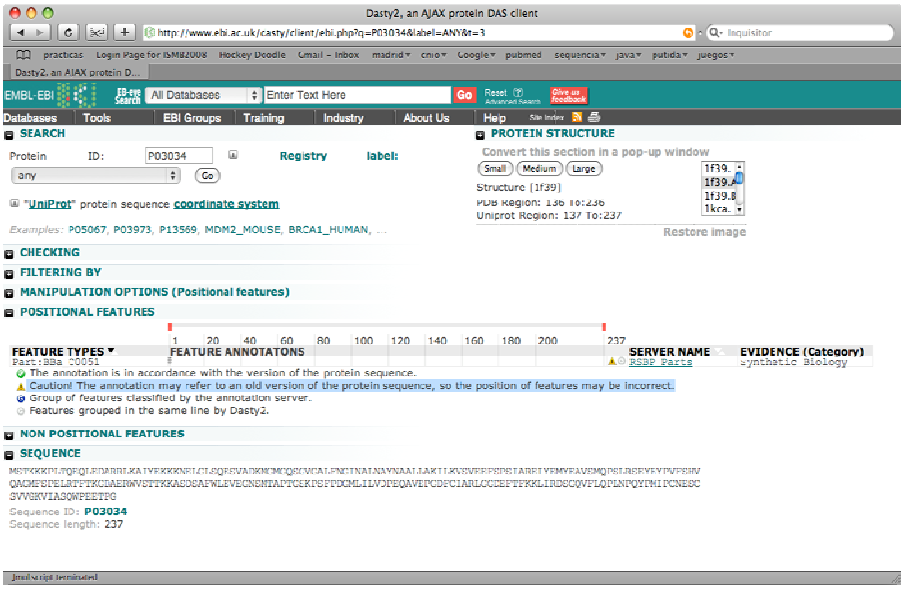
Three experimental DAS servers, with minimal features, were implemented on the MIT site:

- A DAS reference server that basically provide the sequence and IDs of the MIT Registry.
- A DAS annotation server providing annotation about the parts, including subparts, and many other features like, coding sequences, promoters, TF binding sites, terminators, ribosomal binding sites, mutations, etc.
- A DAS annotation server that uses Uniprot as Reference Server which when queried with a Uniprot ID returns if the protein is included in any available part in the Registry.

To demonstrate the feasibility of this approach we have already registered the first DAS servers in the The “DAS registration service” (http://www.dasregistry.org/, Prlic et al. BMC Bioinformatics, 2007).



In this way, the information from the Registry DAS server can now be integrated in other DAS client and shown along other relevant data. One example of this is Dasty, a DAS Client that can e queried with Uniprot IDs. Since Dasty can collect info from DAS servers registered at the DAS registry, it is trivial to incorporate to the Dasty view a track showing the availability of a Biological Part containing the queried Protein.



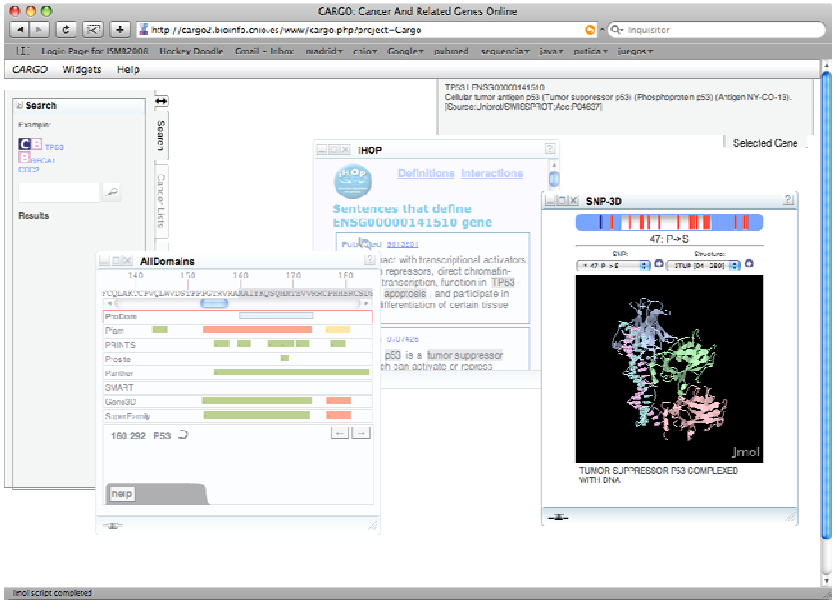
The creation of a more sophisticates and powerful API was discussed in detail, and its practical implementation is being carried out.

The consolidated version of the DAS servers and /or the API site are absolutely essential for the construction of any other bioinformatics tool on top of the MIT Registry structure and / or for the use of the information contained in the Registry in any biological context (including the one described in the next section). This is a clear bottleneck in the project.

5. Initial implementation of tools and methods to visualize / access the information Registry from genomics tools.

Once the information from the Registry will be made accessible following the protocols described above, the next technical steps is its representation in different environments. This exposition is essential to enable the study "parts" in relation with the available biological information, before selecting them for the use of actual Synthetic Biology constructions.

We have developed a system (CARGO; http://cargo2.bioinfo.cnio.es, Cases *et al*., NAR 2007,) for the representation of biological information extracted from different databases and methods. Cargo uses widgets similar to the ones implemented in Google and Macinthos desktop, and currently collects information from more than 20 sources.



The system allows the user to configure the information shown, and the amount of

screen real-state devoted to each data source. For the services provider the system also offers flexibility and each developer is free to design their own visualization framework and to combine different data sources.

The pilot experience for the visualization of Parts in the CARGO environment is the " IGEM parts viewer", which takes advantage of the experimental DAS server described above to display a sketch of the part along with the sequence and provide links to Uniprot when an ID is included in the part annotation.